NIESEMANTYCZNA PRZESTRZEŃ WEKTOROWA

- **podejście surowych częstości słów**: reprezentacja słów w przestrzeni wektorowej
(częstość słowa = liczba wystąpień w dokumencie)
(Raw Word Counts)

- **podejście TF-IDF**: reprezentacja słów w przestrzeni wektorowej
(waga słowa = znormalizowana częstość słowa w dokumencie **x** odwrotna częstość słowa w korpusie)
(Term Frequency-Inverse Document Frequency)

TF-IDF (Term Frequency-Inverse Document Frequency)

- Przezwycięża ograniczenia zwykłej częstości słów, przypisując słowom wagi na podstawie ich częstości w pojedynczym dokumencie oraz w całym zbiorze dokumentów.
- Najwyższe wagi otrzymują słowa, które często pojawiają się w jednym dokumencie, ale rzadko występują w innych dokumentach w korpusie.
- Wynikowa wartość TF-IDF odzwierciedla istotność słowa w badanym korpusie:
słowa z wysoką wartością TF-IDF są uznawane za bardziej istotne, a słowa z niską wartością – za mniej istotne.

Obliczanie TF-IDF

$$TFIDF\ (w, d, D) = TF(w, d) * IDF(w, D)$$

- **TF** mierzy, jak często słowo pojawia się w dokumencie.
Całkowita liczba słów w dokumencie normalizuje TF, gdyż dokumenty mają różną długość:

$$TF(w, d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

- **IDF** mierzy, jak rzadkie lub powszechne jest słowo w całym korpusie dokumentów:

$$IDF(w, D) = \ln(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w})$$

Idea TF-IDF

- Jeśli słowo często występuje tylko w niewielkiej liczbie dokumentów, jego waga (TF-IDF) jest wysoka.
- Słowa powszechnie występujące w tekście (stopwords, np. „i”, „lub”, „the”) otrzymują niskie wagi (TF-IDF), co pozwala je odfiltrować jako mało istotne.

Dlaczego logarytm w IDF?

Zastosowanie logarytmu pozwala na bardziej zrównoważone skalowanie wag i utrzymanie ich w rozsądnych granicach, dzięki czemu TF i IDF mają porównywalny wpływ na wynik końcowy wagi.

Bez zastosowania logarytmu, wartości IDF rosłyby wykładniczo wraz ze wzrostem rzadkości słowa, co prowadziłoby do zdominowania końcowego wyniku przez pojedyncze terminy.

Ograniczenia TF-IDF

- Nie uwzględnia znaczenia semantycznego słów. Na przykład „śmieszny" i „zabawny" są synonimami, ale TF-IDF tego nie rozpozna.
- Może być kosztowne obliczeniowo, gdy korpus zawiera bardzo dużo słów.

Przykład[1]:

Dane wejściowe:

Dokument A: „Jupiter is the largest planet”

Dokument B: „Mars is the fourth planet from the sun”

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

1. **Lista wszystkich słów w korpusie (9 unikalnych słów):**

Jupiter, is, the, largest, planet, Mars, fourth, from, Sun

2. **Obliczenie TF**

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

Dokument A, słowa: jupiter, is, the, largest, planet. Liczba słów: 5
Dokument B, słowa: mars, is, the, fourth, planet, from, the, sun. Liczba słów: 8

Words	TF (for A)	TF (for B)
Jupiter	1/5	0
Is	1/5	1/8
The	1/5	2/8
largest	1/5	0
Planet	1/5	1/8
Mars	0	1/8
Fourth	0	1/8
From	0	1/8
Sun	0	1/8

[1] Na podstawie: https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d/

3. Obliczenie IDF

$$IDF(w, D) = \ln(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w})$$

Liczba dokumentów N = 2

Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

4. Obliczenie TF-IDF

$$TFIDF\ (w, d, D) = TF(w, d) * IDF(w, D)$$

Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

5. Macierz częstości z wagami TF-IDF

Wartość TF-IDF dla danego słowa odzwierciedla jego znaczenie w konkretnym dokumencie, uwzględniając jednocześnie jego częstość występowania w całym zbiorze dokumentów.

Words	TFIDF (A)	TFIDF (B)
Jupiter	0.138	0
Is	0	0
The	0	0
largest	0.138	0
Planet	0	0
Mars	0	0.086
Fourth	0	0.086
From	0	0.086
Sun	0	0.086

Inny przykład: